Mama/Papa, Is this Text for Me?

Md-Rashedur Rahman **Gwénolé Lecorvé** **Jonathan Chevelu**
Univ Rennes, CNRS, IRISA
Lannion, France
first.last@irisa.fr

Nicolas Béchet
Université Bretagne Sud
Vannes, France
first.last@irisa.fr

Aline Étienne **Delphine Battistelli**
Université Paris-Nanterre, CNRS, MoDyCo
Paris, France
first.last@parisnanterre.fr

Abstract

Children have less linguistic skills than adults, which makes it more difficult for them to understand some texts, for instance when browsing the Internet. In this context, we present a novel method which predicts the minimal age from which a text can be understood. This method analyses each sentence of a text using a recurrent neural network, and then aggregates this information to provide the text-level prediction. Different approaches are proposed and compared to baseline models, at sentence and text levels. Experiments are carried out on a corpus of $1,500$ texts and 160K sentences. Our best model, based on LSTMs, outperforms state-of-the-art results and achieves mean absolute errors of 1.86 and 2.28, at sentence and text levels, respectively.

1 Introduction

In recent years, safe Internet for children has gained interest in many research domains (Tomczyk and Kopeckỳ, 2016; Byrne and Burton, 2017; Livingstone, 2019). However, most studies focus on abusive texts containing hate, violence, pornography, etc. (Liu and Forss, 2015; Suvorov et al., 2013). On the contrary, the adequacy of textual contents with the reading and understanding capabilities of children remains yet mainly unresolved in computational linguistics. Hence, this paper propose a new method to predicting this adequacy.

Among the related works, (Schwarm and Ostendorf, 2005) explored the possibility of predicting from which US school grade newspaper articles could be read. This task was modelled as a classification problem among 4 classes using support vector machine fed with word-based n-gram probabilities, as well as lexical and syntactic features. (Islam and Rahman, 2014) has proposed a readability classification method for Bangla news articles for children. This method predicts if a text is either very easy, easy, medium or difficult. More recently, (Blandin et al., 2020) proposed different feed-forward (FF) neural models for age recommendation on texts targeting either children (from 0 to 14) or adults. The authors consider this as a regression task and explore various linguistic features and word embedding features, from which word embedding are shown as the most contributory. Overall, one can notice that these papers either rely on hand-crafted features or on simple models which consider texts as a global object rather than word sequences. This motivates us to further explore with word embeddings only and introduce recurrent neural networks (RNNs).

More broadly, in the field of text readability, inherited from historical approaches like (Kincaid and Chissom, 1975), audiences other than children have been studied, e.g., second language learners (Xia et al., 2016), adults readers (Crossley et al., 2017) or patients interacting with doctors (Balyan et al., 2019). In parallel, text understanding by children is a well known question in psycho-linguistics and cognitive sciences. In particular, key findings have shown the impact of memory (Gathercole, 1999), temporality (Tartas, 2010; Hickmann, 2012), and emotions (Davidson, 2006; Mouw et al., 2019). Related results also exist in the learning to read domain (Frith, 1985).

Proceedings of the 28th International Conference on Computational Linguistics, pages 6296–6301
Barcelona, Spain (Online), December 8-13, 2020

Following the recent trends in NLP, the contribution of this paper is to tackle age prediction as a regression problem using RNNs based on Long Short Term Memories (LSTMs) and fed with pre-trained word embeddings. We propose several variants of this architecture and compare them extensively to naive and FF approaches. We also investigate the difference between predicting age at the sentence and text levels. Let one note that the use of more advanced architectures like transformers (Vaswani et al., 2017) is left for the future, since they are known to require very large amounts of data. The experiments are carried out on a French corpus of around $1,500$ texts of 160K sentences, from encyclopedia, newspapers and fictions for a wide range of different age levels, including adults. We think that this corpus is another interesting aspect of our work, since, compared to others, it is not limited to a specific genre or public.

In the remainder, Section 2 defines with more precision the age prediction task and presents the related data. Then, Section 3 presents the adopted approach and the underlying models. Finally, Section 4 details and discusses the results.

2 Definition of the Problem and Data

In this paper, we consider texts annotated with recommended age ranges $[a, b]$. These age ranges are interpreted as an approximation of the minimal age from which the text can be understood. That is, we define the target minimal age as the mean of the interval, $\frac{a+b}{2}$. Then, to predict this value for a given text, we decide to decompose the problem down to the sentence level, by associating each sentence of a text with the same age range and mean as the whole text. Although this assumption is strong as the complexity of the sentences may vary, it has been shown to be an effective strategy (Blandin et al., 2020).

In practice, we have built a French dataset compiled from encyclopedia, newspapers, and fictions, either dedicated to children or adults. This dataset consists of 1500 texts and abount 160K sentences[1]. Children texts ranges from 0 to 14 years, while adult texts are arbitrary associated with the range $[14, 18]$, and mean 16. Overall, the average of the age range is around 10-14 years where the mean age is 12. The dataset is split into train, dev and test sets at the text level, i.e. all sentences of a given text are kept together in the same set. This partitioning follows the proportions 67/17/16% in terms of sentences, and 70/15/15% in terms of texts. Detailed statistics are provided in Table 1, while distributions over the different ages for each set are given by Figure 1.

3 Approach and Models

Since ages are continuous and sequential values, we consider age prediction as a regression task. It is studied at the sentence and text levels.

[1] The dataset is provided as supplementary material where words are mapped to their respective embedding (see Section 3). Data URL: `https://drive.google.com/file/d/1g7a9VS4G1JIDBjua7HBFzut0_FEJBWCX/view?usp=sharing`

	Train				**Dev**				**Test**			
Genre	**Texts**	**Sent.**	**Age Range**	**Mean**	**Texts**	**Sent.**	**Age Range**	**Mean**	**Texts**	**Sent.**	**Age Range**	**Mean**
Encyclop.	254	40,000	12.53-16.74	14.63	57	10,473	11.90-16.22	14.06	47	7,958	12.64-16.83	14.73
Fiction	397	47,354	9.05-11.84	10.44	93	12,566	7.86-10.9	9.38	78	13,237	8.66-11.82	10.24
Newspaper	391	18,247	9.44-13.85	11.64	74	3,224	9.38-13.78	11.58	96	4,118	9.49-13.86	11.67
Overall	**1,051**	**106,001**	**10.45-14.06**	**12.25**	**225**	**26,334**	**9.67-13.38**	**11.53**	**223**	**25,385**	**10.06-13.74**	**11.90**

Table 1: Summary of the age prediction train, dev and test datasets

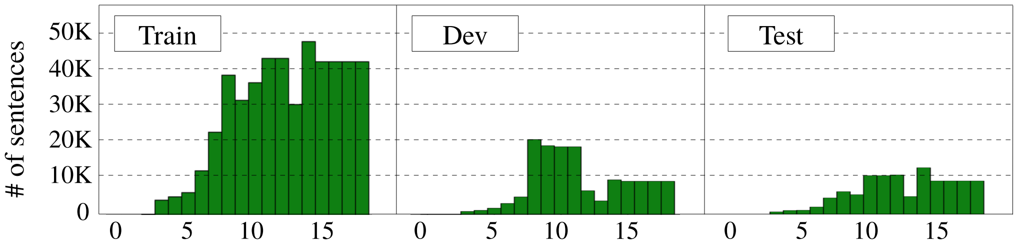

Figure 1: Distribution of the sentences over the different ages in the train, dev and test sets

	Sentence Level		Text Level	
Model	**Dev**	**Test**	**Dev**	**Test**
Naive	3.30	2.90	4.39	4.29
Feed-forward	2.41	2.16	2.59	2.33
LSTM/direct	2.08	**1.86**	**2.31**	**2.28**
LSTM/range	**2.01**	1.89	2.53	2.53
BiLSTM/direct	2.20	2.03	2.49	2.48

Table 2: MAE scores of the different regression models for mean age prediction on dev and test datasets at the sentence and text levels

Our sentence-level models. At the sentence level, the core of the proposed approach is an LSTM-based model (Hochreiter and Schmidhuber, 1997). This kind of RNN is able to learn one-way long-term dependencies of a sequence and is widely used in text classification and time series prediction tasks. In our model, the input is the sequence of words from the input sentence. Each word goes through a projection layer set with pre-trained word embeddings, before entering the LSTM layer. In a first model, the output is directly the real-valued mean age. Alternatively, we also study considering the age range $[a, b]$ as the model's output, before manually deriving the mean. In the remainder, the first option is referred to as ”LSTM/direct”, the second as ”LSTM/range”. We also experiment with the use of a bidirectional LSTM (Schuster and Paliwal, 1997), i.e., the simultaneous use of a forward LSTM and a backward one. The idea is it gets more contextual information on the input data, although the model is more complex (more parameters to be trained). The settings of input and output in this model remain similar to the LSTM/direct model.

Sentence-level baseline models. For comparison, we consider 2 baseline models. The first one consists in a naive approach where sentence-level age prediction is always the mean observed on the training set (12.0). The second is our implementation of the FF model in (Blandin et al., 2020). This model consists in 6 fully-connected layers of 200 units and ReLU activation function, and each input sentence is represented as the average of its word embeddings.

Text-level predictions. Considering either our models or baseline ones for sentence-level predictions, the age prediction for a full text is computed as the average value of the sentence-level predictions.

Training and Parameter Tuning. All models were trained with 50 epochs, using Adam optimizer and the mean squared error as loss. The numbers of LSTM units, batch sizes, dropout, etc. were examined to obtain a robust and stable age prediction model by minimizing the MAE on the development set. In the final experiments, the models are trained with 128 LSTM units, batch size of 256, and dropouts with ratio 0.2 (forward and recurrent ones). The maximum sentence length is set to 100 tokens. Word embeddings are skip-grams trained on FrWaC (Baroni et al., 2009) with dimension 500 and vocabulary size 50K.

4 Experiments

Metrics. As a regression task, results are mainly given in terms mean absolute error (MAE) between the target and predicted ages. To provide a better understanding of the results, the final experiments also evaluate each model as a classifier where each age is a different class. Considering a sentence or text with the reference age range $[a, b]$, a predicted mean age y is considered as *correct* if $y \in [a, b]$, and this correctness is counted as a true positive for each individual whole age part of the age range, as a false negative otherwise. Following this principle, per-class absolute errors are also computed in the final experiments, with a null error if $y \in [a, b]$, and $\min(|y - a|, |y - b|)$ otherwise[2]. For instance, given a reference age range $[5, 7]$, the predictions 5.2 and 7.5 are considered for the classes 5, 6, and 7 as true and false positives, respectively, with absolute errors 0 and 0.5. Doing so, a per-class precision, global accuracy, and per-class MAE can be computed.

[2] This corresponds to the distance of the prediction to the closest bound of the age range.

	Sentence Level				Text Level			
Models	**# Sentences**	**# Correct**	**Avg. Prec.**	**Accuracy**	**# Texts**	**# Correct**	**Avg. Prec.**	**Accuracy**
Naive	25,385	4,749	20.37	18.71	223	24	10.62	10.76
Feed-forward	25,385	13,445	55.85	52.96	223	**119**	**56.32**	**53.36**
LSTM/direct	25,385	**14,839**	**61.33**	**58.46**	223	110	53.48	49.33

Table 3: Classification performances (avg. precision and accuracy) of the models on the test data

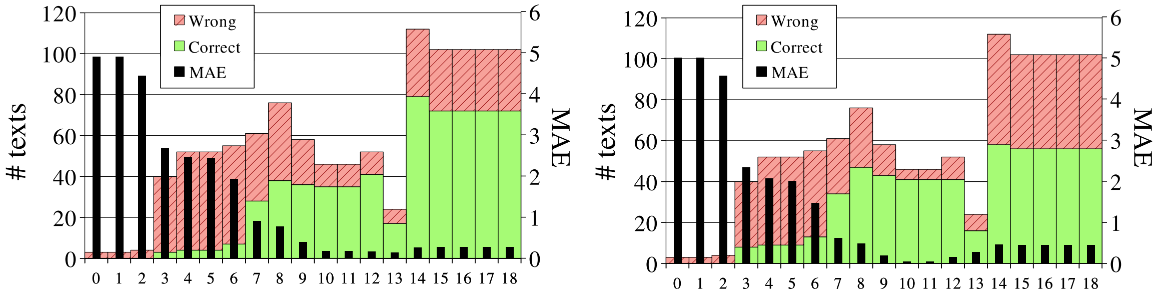

Figure 2: Wrong/correct predictions and MAE per age for the FF (left) and LSTM/direct (right) models

Global results. Table 2 reports MAEs between the target and predicted ages, at the sentence and text levels. First, it appears that all the models perform much better than the naive approach. Then, at the sentence level, the LSTM models significantly outperform the feed-forward model, while no difference between the ”direct” and ”range” approaches appears and the BiLSTM is surpringly doing a bit worse than simple LSTMs. At the text level, it seems that the LSTM/direct is the best RNN model. However, on the test set, it finally does not bring better results than the feed-forward model. Table 3 presents the average precision and accuracy obtained by the naive, FF and LSTM/direct models. Similarly to MAE, the LSTM model achieves the best results at the sentence level. However, this is now the contrary at the text level. While this difference is not very significant given the low number of texts, a deeper investigation discovers that the LSTM/direct model performs worse on the newspaper texts.

Per-class results. Figure 2 details how the feed-forward and LSTM/direct models behave for each age at the text level for the test set. Correct and wrong classifications are given, as well as per-class MAEs. Overall, it appears that both models perform better around the median of the distribution, which seems logical for a machine learning approach. The main difference seems that the feed-forward model is worse than LSTM/direct on very small ages, whereas it is better for adult texts. In complement to previous obversations on genres, these differences probably also contribute to the performance similarity of the FF model on texts inspite of differences at the sentence level. Finally, these results also show that further efforts should be paid on improving predictions for low ages as this is where mistakes would have the strongest impact in a real-life application.

5 Conclusion and perspectives

This paper proposed LSTM models to predict age for sentences and texts. As opposed to the previous related work, these models consider sentences as a sequence of words and do not rely on hand-crafted features. Our best model achieves significantly better scores than the baseline models for the sentence level predictions, confirming the interest of recurrent architectures. However, sentence-level experiments show that this improvement is not propagated at the text level. Hence, in the future, we would like to improve this model for text predictions. To do this, a first perspective is to build more elaborate aggregation techniques of the sentence-level predictions. Then, it would also be interesting to compare with approaches where predictions are directly made at the text-level, without decomposing into sentences. Finally, we would like to explore more advanced recurrent models, like transformers. To do so, an option would be to train a first model using a coarsely annotated corpus, and then fine tuning it on the current corpus. Such imprecise data can be rather easy to collect, for instance, using children-dedicated encyclopedia.

Acknowledgement

We would like to express our gratitude to the French National Research Agency (ANR) as this work has been funded by ANR through the projects TREMoLo (ANR-16-CE23-0019) and TextToKids (AAPG 2019).

References

Renu Balyan, Scott A Crossley, William Brown III, Andrew J Karter, Danielle S McNamara, Jennifer Y Liu, Courtney R Lyles, and Dean Schillinger. 2019. Using natural language processing and machine learning to classify health literacy from secure messages: The eclippse study. *PloS one*, 14(2):e0212488.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3).

Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Étienne. 2020. Age recommendation for texts. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1431–1439.

Jasmina Byrne and Patrick Burton. 2017. Children as internet users: how can evidence better inform policy debate? *Journal of Cyber Policy*, 2(1):39–52.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Denise Davidson. 2006. The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30(3).

Uta Frith. 1985. Beneath the surface of developmental dyslexia. *In K. E. Patterson, J. C. Marshall, & M. Colthear (Eds.),Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*.

Susan Gathercole. 1999. Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, 3, 12.

Maya Hickmann. 2012. Diversité des langues et acquisition du langage: espace et temporalité chez l'enfant. *Langages*, (4).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Zahurul Islam and Rashedur Rahman. 2014. Readability of bangla news articles for children. In *Proceedings of the Pacific Asia Conference on Language, Information and Computing*, pages 309–317.

Robert P. Jr; Rogers Richard L.; Kincaid, J. Peter; Fishburne and Brad S Chissom. 1975. "derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel". *Institute for Simulation and Training*, 02.

Shuhua Liu and Thomas Forss. 2015. Text classification models for web content filtering and online safety. In *Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 961–968. IEEE.

Sonia Livingstone. 2019. Eu kids online. *The international encyclopedia of media literacy*, pages 1–17.

Jolien M Mouw, Linda Van Leijenhorst, Nadira Saab, Marleen S Danel, and Paul van den Broek. 2019. Contributions of emotion understanding to narrative comprehension in children and adults. *European Journal of Developmental Psychology*, 16(1).

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*.

Roman Suvorov, Ilya Sochenkov, and Ilya Tikhomirov. 2013. Method for pornography filtering in the web based on automatic classification and natural language processing. In *Proceedings of the International Conference on Speech and Computer*, pages 233–240. Springer.

Valérie Tartas. 2010. Le développement de notions temporelles par l'enfant. *Développements*, 4.

Łukasz Tomczyk and Kamil Kopeckỳ. 2016. Children and youth safety on the internet: Experiences from czech republic and poland. *Telematics and Informatics*, 33(3):822–833.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*.